Nucleic Acids Research, 2004, Vol. 32, Database issue ***D55–D58***
DOI: 10.1093/nar/gkh059

Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities

Yogita Mantri and Kelly P. Williams*

Department of Biology, Indiana University, 1001 E. Third Street, Bloomington, IN 47405, USA

Received August 15, 2003; Revised September 16, 2003; Accepted September 24, 2003

ABSTRACT

Prokaryotic chromosomes often contain islands, such as temperate phages or pathogenicity islands, delivered by site-specific integrases. Integration usually occurs within a tRNA or tmRNA gene, splitting the gene, yet sequences within the island restore the disrupted gene. The regenerated RNA gene and the displaced fragment of that gene thus mark the endpoints of the island. We applied this principle to search for islands in genomic DNA sequences. Our algorithm generates a list of tRNA and tmRNA genes, uses each as the query for a BLAST search of the starting DNA and removes unlikely hits through a series of filters. A search for islands in 106 whole bacterial genomes produced 143 candidates, with the search itself providing an estimate of three false candidates among these. Preliminary phylogenetic analysis of the associated integrases reduced this set to 89 cases of independently evolved site specificity, which showed strong bias for the tmRNA gene. The website Islander (http://www.indiana.edu/~islander) presents the candidate islands in GenBank-style files and correlates integrase phylogeny with site specificity.

INTRODUCTION

The whole genome sequences for two closely related archaeal or bacterial strains can show near identity throughout most of the genome, yet show dramatic differences in the presence or absence of multigene DNA segments termed islands (1). In addition to sporadic occurrence, islands often exhibit other features of mobile genetic elements that indicate their acquisition by horizontal transfer; many contain a gene encoding an integrase of the tyrosine recombinase family that is responsible for the site-specific positioning of the island. By efficient horizontal delivery of a gene or set of genes that benefits the bacterial host, perhaps promoting pathogenicity, a catabolic pathway or other physiological process, islands are major agents of bacterial evolution. Temperate bacteriophages, which may likewise carry genes beneficial to their bacterial host, can also be considered integrative islands. Site specificity has changed frequently during integrase evolution, which allows diverse islands to accumulate combinatorially in a given host genome. Understanding how site specificity evolves among integrases is therefore a key question in bacterial evolution. We sought to expand the number of known pairs of integrases and the sites they specify, through a bioinformatic search among whole genomes, and in the process have found the endpoints of several previously unrecognized islands. The data from this search, and for other islands from the literature with a known integrase and integration site, are presented at the Islander website (www.indiana.edu/~islander).

SEARCH STRATEGY

By analogy with well-studied integration systems such as that of phage λ (2), it can be presumed that integrative islands exist in circular DNA form prior to integration, and that the integrase catalyzes recombination between a site (*attP*) in the circular pre-island and the target site (*attB*) in the chromosome. Most integrases specify an *attB* that lies within a tRNA or tmRNA gene (tDNA) (3,4). When the island integrates it splits the target tDNA, yet the gene is restored because identical sequence in the *attP* of the island replaces the fragment of the original gene that was displaced. Thus the regenerated tRNA gene and its displaced fragment mark the endpoints of the island. We used this principle in a bioinformatic search for genomic islands that coordinates several pre-existing computer programs. The algorithm proceeds for each genome as follows:

(i) tRNA and tmRNA genes are identified using tRNAscan-SE (5) and BRUCE (6), and the tDNAs with CAT anticodons are sorted into isoleucine, initiator and elongator methionine classes (7).

(ii) Genes for candidate integrases are identified using HMMER with the 'phage integrase' hidden Markov model from Pfam (8), but rejecting those identified as XerC or XerD (housekeeping tyrosine recombinases that do not function as integrases) using Reverse PSI-BLAST (9).

(iii) Each tDNA is used as a query in a BLAST (9) search of the entire genome. The hit and the query gene define the endpoints of a candidate island, which is passed through several filters sequentially.

*To whom correspondence should be addressed. Tel: +1 812 856 5697; Fax: +1 812 855 6705; Email: kellwill@indiana.edu

Table 1. Islands detected in whole bacterial genomes

Bacterial group	Strains	Islands	Islands per strain	Strains with at least one island	Highest per-strain island count
Firmicutes	42	37	0.9	21	5
Escherichia/Shigella/Salmonella	9	45	5.0	9	9
Other γ-proteobacteria	20	28	1.4	13	4
α-Proteobacteria	11	24	2.2	8	6
Other proteobacteria	6	0	0.0	0	0
Other bacteria	18	7	0.4	3	5
Total	106	141		54	

(iv) Candidate islands that do not contain or overlap an integrase open reading frame are rejected.

(v) Remaining candidate islands that entirely contain an integrase open reading frame are rejected if another protein gene of >300 bp overlaps the BLAST hit. This eliminates many false candidates in protein-coding regions, which account for the bulk of a prokaryotic genome. However, integrase genes themselves must be excepted, since some (from the viruses Mx8 and SSV, for example) are known to extend across *attP* (10,11). Currently protein gene coordinates are taken from data (.ptt files) at GenBank (12).

(vi) Remaining candidate islands with the BLAST hit corresponding to a known tRNA gene are rejected. However, it should be noted that a pre-island could in principle contain an entire tRNA gene overlapping the crossover site in *attP*, and therefore generate intact tRNA genes at both of its post-integration endpoints.

(vii) Remaining candidate islands longer than 200 kb are rejected. This cut-off was selected because it would allow detection of all previously known islands except for the 611 kb symbiosis island of *Mesorhizobium loti* (13). In the future we will raise this cut-off since some of our newly detected islands approach it, including a 199 kb island in *Salmonella typhimurium*.

(viii) The tDNA fragments split off by islands are from one end or the other of the original tDNA gene. Remaining candidate islands where the BLAST hit does not extend to one end of the tDNA query are therefore rejected. An exception was made at the 3′ end because certain islands appear to induce upon integration a small deletion at a position in the tDNA fragment 3 bp to the 5′ side of the discriminator position (14). To detect islands with such damaged tDNA fragments, we tolerated BLAST hits that extended only until this deletion site.

(ix) Although the displaced fragment can be from either the 5′ or (more usually) 3′ end of the tRNA gene (15), certain tDNA/fragment configurations are not allowed. Remaining candidate islands with a 5′ fragment downstream of the tDNA, or with a 3′ fragment upstream of the tDNA are rejected.

(x) Cases where multiple remaining candidate islands share the same integrase or the same gene fragment are resolved to single candidate islands. However, multiples sharing the same tDNA are allowed, as tandem arrays, as long as each member of the array would have its own integrase and endpoint.

(xi) Remaining candidate islands with the tRNA gene in the opposite orientation from the BLAST hit are rejected. In principle, such a configuration could produce an invertible DNA segment, but inversion at a tRNA gene has not been reported. This rejection step comes last so that it can serve as a measure of false positives among the final candidates. In genomes without islands, false candidates with the BLAST hit in the opposite orientation to the tRNA gene should be as likely as those in the same orientation.

It should be noted that many integrative islands will be missed by this algorithm, mainly those with an integration site that is not in a tRNA or tmRNA gene. Additionally, several islands are known whose displaced tDNA fragments are too short to be detected by BLAST, at least in its default mode as we currently run it. Vestigial islands with missing or damaged integrase genes would be missed. A small number of temperate phages are known to use an integrase of the serine recombinase family.

SEARCH RESULTS FOR WHOLE BACTERIAL GENOMES

The 106 whole bacterial genomes available at GenBank in July 2003 were searched using the above algorithm, producing 143 final candidates. In the last step of the search algorithm, three had been rejected because the BLAST hit was in the opposite orientation from the tRNA gene, which is an estimate of the number of false positives among the 143 final candidates, as described above. Preliminary inspection allowed the rejection of two of these final candidates, where neighboring islands not themselves integrated into tRNA genes contained clusters of low- or non-scoring tRNA genes. Table 1 shows the phylogenetic breakdown for the remaining 141 candidate islands. Three major groups of bacteria, the Firmicutes (Gram-positives), α-proteobacteria and α-proteobacteria, account for 77% of the strains analyzed, yet contain 95% of the detected islands. The *Escherichia/Shigella/Salmonella* group of γ-proteobacteria average five islands detected per strain, reaching nine in *Escherichia coli* O157:H7 EDL933. For almost half of the bacteria examined, including all the obligate pathogens and endosymbionts, no islands were detected: many of these had no integrase gene.

Partly because some of the genomes analyzed were very closely related, and partly because some clades of closely related integrases have gained a wide host range, multiple islands may represent essentially the same integrase with the same site specificity. Preliminary phylogenetic analysis of the integrases of the 141 islands compressed them into 89 tribes of close relatives with the same site specificity. The catalytic domain sequences of the integrases were aligned using HMMALIGN (8), and pairwise BLOSUM62 distance scores were taken. Tribes were assembled by grouping the integrases

Table 2. Bias in tDNA type specificity among integrase tribes

tDNA	Average no. genes[a]	No. tribes	Bias ratio[b]	P_{bias}[c]
Ala	4.86	0	<0.08	0.0021
Arg	6.70	14	1.71	0.033
Asn	2.65	2	0.62	>0.1
Asp	2.83	0	<0.14	0.029
Cys	1.28	2	1.27	>0.1
Gln	2.83	2	0.58	>0.1
Glu	3.62	1	0.23	0.060
Gly	5.58	7	1.02	>0.1
His	1.28	1	0.64	>0.1
Ile	3.99	6	1.23	>0.1
Leu	6.94	10	1.18	>0.1
Lys	3.46	4	0.94	>0.1
Met	1.96	1	0.42	>0.1
Phe	1.78	4	1.83	>0.1
Pro	3.17	4	1.03	>0.1
Ser	4.63	12	2.12	0.011
Thr	3.94	5	1.04	>0.1
Trp	1.29	0	<0.32	>0.1
Tyr	1.95	1	0.42	>0.1
Val	4.44	2	0.37	0.086
Init	2.69	1	0.30	>0.1
SeC	0.23	1	3.61	>0.1
tmRNA	1.00	9	7.35	0.0000042
Undet[d]	0.41	0	<1.01	>0.1
Total	73.51	89		

[a]Average per host strain, calculated as follows: for each island within a tribe, tDNA types were counted for the host genome, and counts were averaged with weighting according to the summed within-tribe distance scores for the integrase; these values were then averaged for the 89 tribes.
[b]Bias ratio = No. tribes × Total genes/No. genes/Total tribes.
[c]The probability that the observed bias would occur by random assortment, calculated from the binomial distribution.
[d]Anticodon stem–loop distortion renders type undeterminable; probable pseudogenes.

that specified the same tDNA type and had pairwise distance scores of <1.3. One of the tribes had a particularly large and wide-ranging membership, 12 instances of $tRNA^{Ser}$-specificity found throughout the α- and γ-proteobacteria.

The value of sorting integrases into such tribes is that it provides our best view into the evolution of site specificity among integrases, as opposed to the current occupancy of those sites by islands. The distribution of the 89 tribes among 23 types of tDNA is shown in Table 2. Half of the tribes used just four site types: $tDNA^{Arg}$, $tDNA^{Ser}$, $tDNA^{Leu}$ and tmRNA genes, in apparent contrast to the simple expectation that the numbers of tribes using each tDNA type should be proportional to the numbers of those types in the original host genomes. Bias in tDNA type specificity by integrase tribes was calculated, and was significant, either for or against usage, for some types. Bias was highest and extremely significant for the tmRNA gene, and was notable for $tDNA^{Ser}$ and $tDNA^{Arg}$. Bias against the use of $tDNA^{Ala}$ was also highly significant; less so for $tDNA^{Asp}$. The number of tribes will grow as new bacterial genomes and archaeal genomes are included in the analysis, and may change slightly for the current set upon more intensive phylogenetic analysis of the integrases and inspection of the islands on the basis of differences in gene content, codon bias or nucleotide bias from the surrounding host chromosome. However, it can be noted that none of the 40 tribes found in an earlier related study (14) were lost, and that similar tDNA biases were observed in that study, such as the strong bias for tmRNA genes. The new tribes bring coverage to previously unrepresented tDNA types, those specifying Ile, Cys, Gln, Glu, Met and His, and initiator tDNA.

ISLANDER

The islands detected in this search are presented at the Islander website (http:www.indiana.edu/~islander), as GenBank-style files that were abstracted and reindexed from the original whole-genome files (12), together with a variety of summarizing pages. HTML markup highlights the sequences of the target tDNA and the displaced tDNA fragment, as well as the small deletions that occasionally appear in the tDNA fragment. Other pages present integrase alignments and phylogeny. The same information is also presented for several additional integrative islands known from the literature or detected in other ways.

DISCUSSION

As currently implemented, our algorithm does not detect islands integrated into sites outside tRNA genes, and even misses some known islands in tRNA genes if, for example, the displaced gene fragment is too small to be detected by BLAST. However, since the majority of integrative islands are in tRNA genes, we made a rough prior estimate that we would detect half of all islands encoding intact integrases of the tyrosine recombinase family. Among the four complete *E.coli* genomes, where this number has been determined by comparative genome analysis (1,16–18), our success rate was 43% (26 of 61: three of nine in K12, six of 13 in CFT073, eight of 19 in O157:H7 Sakai and nine of 20 in O157:H7 EDL933). In principle, the algorithm could be modified to detect more of the islands in tRNA genes that are now missed, or to search for islands in other types of integration site. Many genes for small RNAs unrelated to tRNA have recently been identified in *E.coli*, and one of these is the integration site for phage P2 (19,20). We made a preliminary attempt to find islands in any of 43 small non-tRNA RNA genes in the whole genomes of enterobacteria, but did not detect any. In any case, a collection of only the integration sites in tDNAs does represent a large number of all sites, and, moreover, the conserved sequences and patterns of tDNAs provide a standardized format that is useful in aligning the sites for fine comparison.

It is hoped that the database will facilitate research into the evolution of site specificity among integrases, and also serve as a useful resource for microbiologists studying bacteriophages or other genetic functions provided by islands.

REFERENCES

1. Perna,N.T., Plunkett,G., Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
2. Campbell,A.M. (1962) Episomes. *Adv. Genet.*, **11**, 101–145.
3. Reiter,W.D., Palm,P. and Yeats,S. (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res.*, **17**, 1907–1914.
4. Williams,K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.*, **30**, 866–875.

5. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
6. Laslett,D., Canback,B. and Andersson,S. (2002) BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3449–3453.
7. Marck,C. and Grosjean,H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, **8**, 1189–1232.
8. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Tojo,N., Sanmiya,K., Sugawara,H., Inouye,S. and Komano,T. (1996) Integration of bacteriophage Mx8 into the *Myxococcus xanthus* chromosome causes a structural alteration at the C-terminal region of the IntP protein. *J. Bacteriol.*, **178**, 4004–4011.
11. Peng,X., Holz,I., Zillig,W., Garrett,R.A. and She,Q. (2000) Evolution of the family of pRN plasmids and their integrase-mediated insertion into the chromosome of the crenarchaeon *Sulfolobus solfataricus*. *J. Mol. Biol.*, **303**, 449–454.
12. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
13. Sullivan,J.T. and Ronson,C.W. (1998) Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc. Natl Acad. Sci. USA*, **95**, 5145–5149.
14. Williams,K.P. (2003) Traffic at the tmRNA gene. *J. Bacteriol.*, **185**, 1059–1070.
15. Zhao,S. and Williams,K.P. (2002) Integrative genetic element that reverses the usual target gene orientation. *J. Bacteriol.*, **184**, 859–860.
16. Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
17. Welch,R.A., Burland,V., Plunkett,G., Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
18. Ohnishi,M., Kurokawa,K. and Hayashi,T. (2001) Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol.*, **9**, 481–485.
19. Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1813–1820.
20. Wassarman,K.M., Repoila,F., Rosenow,C., Storz,G. and Gottesman,S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, **15**, 1637–1651.